Biostatistics (2012), **0**, 0, *pp.* 1–25

doi:10.1093/biostatistics/WANG˙DAI˙FU˙biostats˙revision2˙Aug19

Accelerated failure time models for censored survival data under referral bias

Huan Wang and Hongsheng Dai*, University of Brighton

and Bo Fu, University of Manchester

Summary

The estimation of progression to liver cirrhosis and identifying its risk factors are often of epidemiological interests in hepatitis C natural history study. In most hepatitis C cohort studies, patients were usually recruited to the cohort with referral bias because clinically the patients with more rapid disease progression were preferentially referred to liver clinics. A pair of correlated event times may be observed for each patient, time to development of cirrhosis and time to referral to a cohort. This paper considers accelerated failure time models to study the effects of covariates on progression to cirrhosis. A new non-parametric estimator is proposed to handle a flexible bivariate distribution of the cirrhosis and referral times and to take the referral bias into account. The asymptotic normality of the proposed estimator is also provided. Numerical studies show that the coefficient estimator and its covariance function estimator perform well.

Key words: Accelerated failure time model; Bivariate survival function; Correlated failure times; Censoring; Truncation; Survival analysis.

1. Introduction

The motivation for this paper arose from a hepatitis C cohort study in Fu *et al.* (2007), where the epidemiological interest is to study progression to liver cirrhosis in hepatitis C virus (HCV)-infected patients. Two event times of interest are time from HCV infection to the development of cirrhosis, denoted by T, and time from HCV infection to referral to the clinic cohort, denoted by R. Clinically the patients with more rapid disease progression are preferentially referred to liver clinics or that referral is increasingly likely the closer a patient is to developing cirrhosis (Fu *et al.*, 2007). If so, the conventional analysis based on liver clinic cohorts will lead to biased estimate of progression rate among the HCV patient community (Freeman *et al.*, 2001; Fu *et al.*, 2007). This is usually called referral bias in epidemiology (Dore *et al.*, 2002). Statistically, the bivariate event times (R, T) are obviously correlated, and the patients were included to the cohort with referral bias because only patients who referred to the clinic before the end of study recruitment at time L were included ($R \leqslant L$), i.e. R is right truncated by L. Under truncation a patient will not be observed if the referral occurs after the end of recruitment (no information on this subject is available if $R > L$). The cirrhosis time T is subject to censoring at time C, for example at the last diagnosis follow-up. This is an example of bivariate survival data with both censoring and truncation.

In this paper, we consider a regression model to study the effects of risk factors on the disease progression to cirrhosis, such as age at infection, HIV-coinfection, and alcohol intake. The challenge here is how to take the referral bias into account. Conventional univariate survival analysis, which ignores the referral bias, has been seen to produce seriously biased results in estimating the progression rate to cirrhosis within 20-30 years; see Fu *et al.* (2007). To obtain unbiased estimate of the effects of covariates on incubation time from infection to cirrhosis, we

start with a commonly used model, the accelerated failure time (AFT) model, which is given by

$$\log T_i = \boldsymbol{W_i}\boldsymbol{\beta} + \varepsilon_i, \tag{1.1}$$

where T_i denotes the time from HCV infection to cirrhosis for patient i, $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{ip})$ is the covariate vector and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{tr}$ is the regression coefficient vector. Here we assume that the error terms $\{\varepsilon_i, i = 1, \cdots, n\}$ are i.i.d. having mean 0 and an unknown distribution $F_\varepsilon(\cdot)$ and are independent of the covariates $\boldsymbol{W}$. We also assume that (L, C) is independent of the covariate $\boldsymbol{W}$ and is independent of the time pair (R, T). When T_i is only subject to random right censoring at C_i, i.e. we observe $(X_i, \boldsymbol{W}_i, \delta_i)$ where $X_i = \min(T_i, C_i)$, $\delta_i = I[T_i \leqslant C_i]$, model (1.1) has been well studied (Miller, 1976; Miller and Halpern, 1982; Koul, Susarla and Van, 1981; Buckley and James, 1979; Jin *et al.*, 2006). A comparison study is given in Bao *et al.* (2007). When the response T_i itself is also subject to truncation, model (1.1) can also be solved using the methods in Gross and Huber-Carol (1992), Gross and Lai (1996) and He and Yang (2003).

For the hepatitis C clinic cohort study mentioned above, T_i is subject to censoring but T_i itself is not subject to truncation. However, the observed sample of T_i is biased due to the truncation on the referral time R_i. Existing methods in survival data analysis are not readily available for model (1.1) if referral bias is taken into account. In this paper we are interested in obtaining an unbiased estimate of $\boldsymbol{\beta}$ by taking account of the correlation between T and R. We will introduce a nonparametric method for model (1) which does not assume a particular distribution of ε and does not specify a joint distribution between T and R. The developed method is a new, flexible and important candidate for handling the referral bias in HCV natural history study based on clinic cohort data. Moreover, our method can be directly applied to regression analysis of bivariate survival data with random censoring and random truncation and can be generally extended to a general class of bivariate regression models with different types of truncation and censoring; see discussion in Dai and Fu (2012).

This paper is organized as follows. We begin in Section 2 with a description of the statistical model and introduce an estimating procedure for regression coefficients $\boldsymbol{\beta}$. The new procedure takes referral bias into account using a bivariate survival function estimator based on a polar coordinate transformation. The large sample properties of the estimator for $\boldsymbol{\beta}$ is given in Section 3. Section 4 gives simulation studies and data analysis which demonstrate that the proposed estimator for $\boldsymbol{\beta}$ performs well. Section 5 gives a short discussion.

2. Notations and estimation procedure

2.1 *Notations and estimating equations*

For simplicity we use (R_i, T_i) to denote the pair of log-event times for the ith subject, where R_i is the logarithm of time from hepatitis C virus infection to referral to the cohort, and T_i is the logarithm of time from infection to the development of cirrhosis. The value R_i, is randomly right-truncated at L_i, which is the logarithm of recruitment time. The value T_i, is subject to random right censoring at C_i, which is the logarithm of the censoring time. If $R_i > L_i$ then we cannot get any information about this patient. If $R_i \leqslant L_i$, then we can observe $\{R_i, L_i, X_i, \delta_i, \boldsymbol{W}_i, \}$, where $\delta_i = I[T_i \leqslant C_i]$, $X_i = \min\{T_i, C_i\}$ and $\boldsymbol{W}_i$ is the covariate vector. We usually denote the observed data as $\{R_i^*, L_i^*, X_i^*, \delta_i^*, \boldsymbol{W}_i^*\}$, $i = 1, \cdots, n$. We assume that $(R, T, \boldsymbol{W})$ and (L, C) are independent throughout this paper.

Let $G(t_1, t_2) = P(L_i > t_1, C_i > t_2)$ be the continuous bivariate survival function for (L_i, C_i) and $\bar{F}(t_1, t_2) = P(R_i > t_1, T_i > t_2)$ be the continuous joint survival function for (R_i, T_i). Let $F(t_1, t_2, \boldsymbol{w})$ be the joint distribution function for R_i, T_i and $\boldsymbol{W}_i$. Let $F^*(x_1, x_2, \boldsymbol{w}) = P(R_i^* \leqslant x_1, X_i^* \leqslant x_2, \boldsymbol{W}_i^* \leqslant \boldsymbol{w}, \delta_i^* = 1)$ be the joint cumulative distribution function for the observed vector $(R_i^*, X_i^*, \boldsymbol{W}_i^*, \delta_i^* = 1)$.

Define the boundaries of support for the density of $\bar{F}$ as

$$\boldsymbol{a}_{\bar{F}} = \inf\{(t_1, t_2) : \bar{F}(t_1, t_2) < 1\},\ \boldsymbol{b}_{\bar{F}} = \sup\{(t_1, t_2) : \bar{F}(t_1, t_2) > 0\} \tag{2.2}$$

and define $\boldsymbol{a}_G$, $\boldsymbol{b}_G$ similarly. For simplicity we assume that the $\boldsymbol{a}_{\bar{F}}$ are the two coordinated axis, which means $R \geqslant 0, T \geqslant 0$. We assume that the following conditions hold.

Condition 2.1 (i) For $\boldsymbol{s} = (s_1, s_2) \in \boldsymbol{b}_{\bar{F}}$ and $\boldsymbol{t} = (t_1, t_2) \in \boldsymbol{b}_G$, we have $\sqrt{s_1^2 + s_2^2} < \sqrt{t_1^2 + t_2^2}$ given $s_2/s_1 = t_2/t_1$. (ii) For $\boldsymbol{s} \in \boldsymbol{a}_{\bar{F}}$ and $\boldsymbol{t} \in \boldsymbol{a}_G$, we have $\sqrt{s_1^2 + s_2^2} < \sqrt{t_1^2 + t_2^2}$ given $s_2/s_1 = t_2/t_1$.

Here $\boldsymbol{s} \prec \boldsymbol{t}$ means that $s_1 < t_1, s_2 \leqslant t_2$ or $s_1 \leqslant t_1, s_2 < t_2$. These conditions are realistic in our hepatitis C study, which is explained in appendix. Under (i) of Condition 2.1 for all $\boldsymbol{t} = (t_1, t_2)$, $\boldsymbol{0} \leqslant \boldsymbol{t} \leqslant \boldsymbol{b}_{\bar{F}}$, we have $G(t_1, t_2) > 0$. Therefore we have the following relation,

$$F(t_1, t_2, \boldsymbol{w}) = \gamma \cdot \int_{s_1 \leqslant t_1} \int_{s_2 \leqslant t_2} \int_{\boldsymbol{u} \leqslant \boldsymbol{w}} \frac{1}{G(s_1-, s_2-)} F^*(ds_1, ds_2, d\boldsymbol{u}) \tag{2.3}$$

where $\gamma = P(R_i \leqslant L_i)$ is the truncation probability. Note that (i) of Condition 2.1 guarantees that the function $\bar{F}(t_1, t_2)$ is identifiable in the whole support region from $\boldsymbol{0}$ to $\boldsymbol{b}_{\bar{F}}$.

The following lemma gives an unbiased estimating equation for $\boldsymbol{\beta}$, provided $G(t_1, t_2)$ is known.

Lemma 2.1 Parameter $\boldsymbol{\beta}$ satisfies

$$\boldsymbol{q} = \boldsymbol{\Gamma}\boldsymbol{\beta} \tag{2.4}$$

where $\boldsymbol{\Gamma} = (\varsigma_{jk})_{j,k=1}^{p}$, $\varsigma_{jk} = E(W_{ij}W_{ik}) = \int w_j w_k F(dt_1, dt_2, d\boldsymbol{w})$ and $\boldsymbol{q} = (\varsigma_{01}, \cdots, \varsigma_{0p})^{tr}$, $\varsigma_{0k} = E(W_{ik}T_i) = \int t_2 w_k F(dt_1, dt_2, d\boldsymbol{w})$.

If G is known, ς_{0k} and ς_{jk}, can be estimated respectively by the unbiased estimates

$$\hat{\varsigma}_{0k}(G) = \frac{\hat{\gamma}}{n}\sum_{i=1}^{n}\frac{X_i^* W_{ik}^* \delta_i^*}{G(R_i^*-, X_i^*-)},\ \hat{\varsigma}_{jk}(G) = \frac{\hat{\gamma}}{n}\sum_{i=1}^{n}\frac{W_{ij}^* W_{ik}^* \delta_i^*}{G(R_i^*-, X_i^*-)}, \tag{2.5}$$

where $\hat{\gamma}$ is the truncation probability estimate. Then we can obtain a consistent estimate

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\Gamma}^{-1}(\hat{G})\boldsymbol{q}(\hat{G}), \tag{2.6}$$

if $\hat{G}(t_1, t_2)$ is a consistent estimate for the function $G(t_1, t_2)$. □

Dai and Fu (2012) provide a method to estimate the truncation probability γ. We, however, do not need the truncation probability $\hat{\gamma}$ when estimating $\boldsymbol{\beta}$, since $\hat{\gamma}$ is cancelled out in (2.6). The above method is actually the weighted least squares (WLS) method. See He and Wong (2003) for more details and the idea of proving Lemma 2.1 can also be found therein.

2.2 *Estimation of the bivariate survival function G*

Part (ii) of Condition 2.1 guarantees that for any observed L there exists R such that $R < L$, which means all values of L can be possibly observed. So under Condition 2.1 we know that G is identifiable in the region from $\mathbf{0}$ to $\boldsymbol{b}_{\bar{F}}$. There is a vast literature about nonparametric estimation of bivariate survival functions $G(t_1, t_2)$ under right censoring (Campbell, 1981; Tsai *et al.*, 1986; Burke, 1988; Dabrowska, 1988; Lin and Ying, 1993; van der Laan, 1996a; Akritas and Keilegom, 2003; Prentice *et.al.*, 2004; Dai and Bao, 2009). Bivariate distribution estimation under truncation was also well studied; see for example, Gurler (1996, 1997) for applications when a single component of the bivariate data is subject to truncation; van der Laan (1996b) and Huang *et al.* (2001) for data under bivariate truncation. Gijbels and Gurler (1998) proposed an interesting nonparametric estimator for bivariate data where a single component is subject to both censoring and truncation and the other component is fully observed.

The above methods are not readily available here for estimating $G(t_1, t_2) = P(L > t_1, C > t_2)$ as L is subject to left-truncation by R, C is subject to right-censoring by T and (R, T) are correlated. For bivariate survival data where both components are under censoring and truncation, Shen (2006) proposed an inverse probability weighted (IPW) approach to estimate the joint survival function. His method requires an iterative algorithm to calculate the distribution estimate, which is computationally heavy and might be impractical for data with a large sample size. Recently Dai and Fu (2012) proposed an estimator for such paired survival times under both truncation and censoring. Their method is based on a polar coordinate transformation,

$$G(t_1, t_2) = P(L > t_1, C > t_2) = P(Z(\alpha) > z) := G(z; \alpha),$$

where $\alpha = t_2/t_1$, $z = \sqrt{{t_1}^2 + {t_2}^2}$ and $Z(\alpha) = \min\{L\sqrt{1+\alpha^2}, C\sqrt{1+\alpha^{-2}}\}$. Dai and Fu (2012) proposed a consistent estimator for the transformed survival function $G(z; \alpha)$. This transformation method does not require a heavy load of computation. In this section, we will use their method to estimate the bivariate survival function G.

In practice, due to censoring and truncation, the values of (L, C) may not be obtained. Thus $Z(\alpha)$ may not be available. However, we have the observed transformed data:

$$\tilde{Z}^*(\alpha) = \min\{\tilde{L}^*, \tilde{X}^*\}, \Delta^*(\alpha) = I[\tilde{L}^* \leqslant \tilde{X}^*] + (1-\delta^*)I[\tilde{L}^* > \tilde{X}^*], V^*(\alpha) = R^*\sqrt{1+\alpha^2}, \quad (2.7)$$

where $\tilde{L}^* = L^*\sqrt{1+\alpha^2}$, $\tilde{X}^* = X^*\sqrt{1+\alpha^{-2}}$ and L^*, R^*, X^*, δ^* are the observed data. Note that such a transformation introduces artificial censoring and truncation. For example $\Delta^*(\alpha) = 1$ implies that $\tilde{Z}^*(\alpha)$ is an observed value for $Z(\alpha)$ and $\Delta^*(\alpha) = 0$ implies censoring. Truncation information is given by $V^*(\alpha)$. However, the truncation condition is not equivalent to $\tilde{Z}(\alpha) \geqslant V(\alpha)$. More detailed discussion can be found in Dai and Fu (2012). Then based on the transformed observations in (2.7) we can estimate $G(z; \alpha)$ using the following lemma, the proof of which can

be found in Dai and Fu (2012).

LEMMA 2.2 For fixed α, the hazard rate function of $Z(\alpha)$ is denoted by $\Lambda(dz;\alpha) = -\frac{G(dz;\alpha)}{G(z-;\alpha)}$. Then we have

$$\Lambda(dz;\alpha) = \frac{P(\tilde{Z}^*(\alpha) \in dz, z > V^*(\alpha), \Delta^*(\alpha) = 1)}{P(\tilde{Z}^*(\alpha) \geqslant z > V^*(\alpha))},$$

where $\tilde{Z}^*(\alpha) \in dz$ denotes $z \leqslant \tilde{Z}^*(\alpha) < z + dz$. □

We define $N(ds;\alpha) = n^{-1}\sum_{i=1}^n N_i(ds;\alpha) = n^{-1}\sum_{i=1}^n I[\tilde{Z}_i^*(\alpha) \in ds, s > V_i^*(\alpha), \Delta_i^*(\alpha) = 1]$, $H_{(n)}(s;\alpha) = n^{-1}\sum_{i=1}^n H_i(s;\alpha) = n^{-1}\sum_{i=1}^n I[\tilde{Z}_i^*(\alpha) > s \geqslant V_i^*(\alpha)]/n$ and $H_{(n)}(t_1,t_2) = n^{-1}\sum_{i=1}^n H_i(t_1,t_2) = n^{-1}\sum_{i=1}^n I\left[L_i^* > t_1 \geqslant R_i^*, X_i^* > t_2\right]$. Note that $H_{(n)}(t_1,t_2) = H_{(n)}(z;\alpha)$ and $H_i(t_1,t_2) = H_i(z;\alpha)$. Lemma 2.2 implies that an estimator for $\Lambda(dz;\alpha)$ is $\hat{\Lambda}(dz;\alpha) = N(dz;\alpha)/H_{(n)}(z-;\alpha)$. Then the product-limit estimator for $G(z;\alpha)$ is

$$\hat{G}(z;\alpha) = \prod_{s\leqslant z}^n \left\{1 - \frac{\Delta N(s;\alpha)}{H_{(n)}(s-;\alpha)}\right\}, \tag{2.8}$$

where $\Delta N(s;\alpha) = N(s;\alpha) - N(s-;\alpha)$.

Define $M_j(ds;\alpha) = N_j(ds;\alpha) - H_j(s-;\alpha)\Lambda(ds;\alpha)$, $H(s;\alpha) = E[H_i(s;\alpha)] = G(s_1,s_2)P[s_1 > R, T > s_2]/\gamma$, and define τ_α as $\tau_\alpha = \sqrt{t_1^2+t_2^2}$ with $(t_1,t_2) \in \boldsymbol{b}_{\bar{F}}, t_2/t_1 = \alpha$. So $[0,\tau_\alpha]\times[0,\infty]$ can be viewed as the range for $(\tilde{Z}_i^*(\alpha_i^*), \alpha_i^*)$. Following the ideas in Dai and Fu (2012) we can show the following asymptotic results for $\hat{G}$,

$$\sqrt{n}\left[\hat{G}(z;\alpha) - G(z;\alpha)\right] = -\frac{1}{\sqrt{n}}G(z;\alpha)\sum_{j=1}^n \int_{s\leqslant z} \frac{1}{H(s-;\alpha)} M_j(ds;\alpha) + r_n(z;\alpha), \tag{2.9}$$

where $r_n(z;\alpha)$ is a term such that $\sup_{z\in[0,\tau_\alpha],\alpha\in[0,\infty]} E[r_n(z;\alpha)]^2 \to 0$.

Since $G(z;\alpha) = G(t_1,t_2)$, $\hat{G}(z;\alpha)$ is also an estimator for the bivariate survival function $G(t_1,t_2)$. By substituting this $\hat{G}$ into (2.6), we can get the estimator $\hat{\boldsymbol{\beta}}(\hat{G})$.

3. Asymptotic normality for $\hat{\boldsymbol{\beta}}$

The estimate in (2.6) is equivalent to solving the estimating equation

$$\boldsymbol{Q}\left(\boldsymbol{\beta};\hat{G}\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{G}(R_i^*-,X_i^*-)}(X_i^*-\boldsymbol{W}_i^*\boldsymbol{\beta})\boldsymbol{W}_i^{*tr}=0, \tag{3.10}$$

since $\boldsymbol{Q}\left(\hat{\boldsymbol{\beta}};\hat{G}\right)=\boldsymbol{\Gamma}(\hat{G})\hat{\boldsymbol{\beta}}-\boldsymbol{q}(\hat{G})=\mathbf{0}$. The following theorem provides the results of asymptotic normality.

Theorem 3.1 Let

$$\boldsymbol{\eta}_i=\delta_i^*(X_i^*-\boldsymbol{W}_i^*\boldsymbol{\beta})\boldsymbol{W}_i^{*tr},\ \ \xi_{ji}=\int_{s\leqslant\tilde{Z}_i^*(\alpha_i)}\frac{1}{H(s-;\alpha_i)}M_j(ds;\alpha_i) \tag{3.11}$$

where $\alpha_i=X_i^*/R_i^*$. Let $\boldsymbol{Q}^{'}(\boldsymbol{\beta};G)=\partial\boldsymbol{Q}(\boldsymbol{\beta};G)/\partial\boldsymbol{\beta}$ and $\mathcal{D}_k=\{R_k^*,L_k^*,X_k^*,\delta_k^*,\boldsymbol{W}_k^*\}$ denote the observed information of patient k. Then we have that $\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\right)\sim N(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}=\left[\boldsymbol{Q}^{'}(\boldsymbol{\beta};G)\right]^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Q}}\left\{\left[\boldsymbol{Q}^{'}(\boldsymbol{\beta};G)\right]^{-1}\right\}^{tr}$. The matrix $\boldsymbol{\Sigma}_{\boldsymbol{Q}}$ is given by

$$\boldsymbol{\Sigma}_{\boldsymbol{Q}}=\mathrm{Var}\left[\frac{\boldsymbol{\eta}_k}{G(R_k^*,X_k^*)}+\boldsymbol{\Phi}(\mathcal{D}_k)\right],\ \boldsymbol{\Phi}(\mathcal{D}_k)=E\left\{G^{-1}(R_i^*,X_i^*)\xi_{ki}\boldsymbol{\eta}_i\,|\mathcal{D}_k\right\},\ i\neq k. \tag{3.12}$$

The proof of the theorem can be found in appendix. Use $\hat{\boldsymbol{\eta}}_i,\hat{\boldsymbol{\Phi}}(\mathcal{D}_i)$ to denote the estimates for $\boldsymbol{\eta}_i,\boldsymbol{\Phi}(\mathcal{D}_i)$ (replacing G, $\boldsymbol{\beta}$ with $\hat{G}$, $\hat{\boldsymbol{\beta}}$ respectively in $\boldsymbol{\eta}_i,\boldsymbol{\Phi}(\mathcal{D}_i)$). Then an estimate of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}=\left[\boldsymbol{Q}^{'}(\hat{\boldsymbol{\beta}};\hat{G})\right]^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Q}}\left\{\left[\boldsymbol{Q}^{'}(\hat{\boldsymbol{\beta}};\hat{G})\right]^{-1}\right\}^{tr}, \tag{3.13}$$

where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Q}}=\widehat{var}\left[\frac{\hat{\boldsymbol{\eta}}_i}{\hat{G}(R_i^*,X_i^*)}+\hat{\boldsymbol{\Phi}}(\mathcal{D}_i)\right]$ is the sample covariance matrix based on $\dfrac{\hat{\boldsymbol{\eta}}_i}{\hat{G}(R_i^*,X_i^*)}+\hat{\boldsymbol{\Phi}}(\mathcal{D}_i)$, $i=1,\cdots,n$.

4. Simulation studies and data analysis

4.1 *Simulation studies*

In this section, we study the properties of our proposed estimator for $\boldsymbol{\beta}$ via a set of 500 simulations. Truncation times L and censoring times C are generated respectively from $C = a\nu_1 + b\nu_2,\ L = c\nu_1 + d\nu_2 + U[0,1]$ where ν_1 and ν_2 are exponentially distributed with unit mean. We can change the values of (a, b, c, d) to adjust censoring/truncation probabilities and correlations of L and C. Pairs of survival times are generated from the following model to mimic the data analysis. The logarithm of survival time follows $T = \boldsymbol{W}\boldsymbol{\beta}^* + \varepsilon$, where the true value $\boldsymbol{\beta}^* = (3.7, -0.05, -0.3, -0.1)^T$ and the covariate $\boldsymbol{W}$ is defined as: the intercept $W_1 = 1$ and the predictors $W_2 \sim U[20, 30]$, W_3 ∼Bernoulli(0.5) and W_4 ∼Bernoulli(0.5). The time R has mean $ER = 1.35$ and (R, T) are correlated through a joint distribution of $(R - ER, \varepsilon)$. Once we simulate $(R - ER, \varepsilon)$ we then have the simulated values for R and $T = \boldsymbol{W}\boldsymbol{\beta}^* + \varepsilon$. We here consider two types of joint distribution functions of $(R - ER, \varepsilon)$.

Scenario 1: The two error terms $R - ER$ and ε are generated as follows, $R - ER = 1.0\upsilon_1 + 0.5\upsilon_2,\ \ \varepsilon = 0.4\upsilon_1 + 0.35\upsilon_2$, where υ_1 and υ_2 are two uniform random variables from $\mathbf{U}[-0.5, 0.5]$. The linear combinations above can make $R - ER$ and ε correlated. The simulation results are presented in Table 1. We choose different values for a, b, c, d to achieve different censoring and truncation percentages.

Table 1 is about here.

The simulation results indicate that the biases are 0.001 for all parameters, which is very small, when sample size is large ($n = 200$), censoring percentage is low (about 20%) and truncation probability is high (about 0.85). In this case, the mean standard deviation estimate and standard deviation for Monte Carlo estimates are very close. Even when the data are severely biased (80% censoring and 0.15 truncation probability), the biases for the predictor parameters $(\beta_2, \beta_3, \beta_4)$ are still very small, $(0.001, 0.010, 0.004)$ respectively, although the bias for intercept β_1 is larger.

Therefore we can conclude that the proposed estimators work well for large sample sizes.

Note that when sample size is $n = 100$, censoring probability is about 80% and truncation probability is about 0.15, the estimate for the intercept has the largest biases 0.064 and the estimates for other parameters are still good. In this case the standard error based on Monte Carlo simulations $\hat{s}_\beta$ and the mean of estimated standard errors $\hat{\sigma}_\beta$ are not close. This is not surprising as the observed sample is severely biased (truncation probability is only about 0.15) and among the observed samples 80% are censored, i.e. only about 20 observations are fully observed.

Scenario 2: Instead of generating the error terms from linear combinations, we generate $R - ER$ and ε as follows. First generate (ω_1, ω_2) from the well-known bivariate parametric model in Clayton (1978), which has joint survival function $S_\epsilon(s_1, s_2) = (S_{\epsilon 1}(s_1)^{-\phi} + S_{\epsilon 2}(s_2)^{-\phi} - 1)^{-1/\phi}$. We take $\phi = 4$ (see for example Prentice *et.al.* (2004)). The marginal survival functions $S_i(s_i)$ are from unit exponential distribution truncated at 1 and its mean is $E = (1 - 2e^{-1})/(1 - e^{-1})$. Then we let $R = ER + \omega_1 - E$ and $\varepsilon = \omega_2 - E$. We can also choose different values for a, b, c, d to achieve different censoring and truncation percentages. The simulation results are also summarized in Table 2. We have similar findings and conclusions as for Scenario 1.

Table 2 is about here.

The boundary conditions are satisfied in both scenarios in Section 4.1. Note that (R, T) has a finite upper boundary for their density support but (L, C) are from linear combinations of exponential distributions and can take values up to ∞. This guarantees that there is always a positive probability to observe all possible values for (R, T), which makes the model identifiable.

On the contrary, if the boundary condition is not satisfied, the estimates could be biased. For example, we consider **Scenario 3** by drawing L and C from $C = a\nu_1 + b\nu_2, \;\; L = c\nu_1 + d\nu_2$, where ν_1 and ν_2 are from unit exponential distribution truncated at 2. The two error terms $R - ER$ and ε are generated as $R - ER = 0.5\upsilon_1 + 0.5\upsilon_2, \;\; \varepsilon = 0.4\upsilon_1 + 0.4\upsilon_2$, where $ER = 0.5$, υ_1 and υ_2

are from Normal$(0, 1.0)$. In this scenario, the upper boundary for the support of the density of L is $2(c+d)+1$. For all (R, T) with $R > 2(c+d)+1$ can never be observed. Also larger value of T will always be censored. Therefore, if $c+d$ is small the estimates will be severely biased. This is also shown in Table 2.

From the results we can see that the intercept is estimated much smaller than the true value and the predictor parameter estimates are much less significant than that in Scenarios 1 and 2. This is because large values of R can never be observed and large values of T will always be censored. This will result in a much smaller intercept estimate and less significant predictor parameter estimates.

4.2 *Data analysis*

We illustrate the proposed method with the Edinburgh hepatitis C data in Fu *et al.* (2007). The data set consists of 387 HCV-infected individuals recruited from Edinburgh Royal Infirmary's liver clinic by the end of 1999, for whom mean age at HCV-infection is around 22 years. These patients were studied retrospectively and followed prospectively for the development of HCV-related cirrhosis. HCV patients usually experience no symptoms or mild symptoms in the early stages and are often referred to hospital shortly before they develop cirrhosis or complications. Among these individuals, there is no cirrhosis event occurred prior to their referral time and 63 (16%) developed cirrhosis during follow-up. The median during time from HCV infection to referral is 17.1 years and the median follow-up time from referral to cirrhosis or censoring is 2.4 years. The demographic details are shown in Table 3. An individual's information is only available if he or she was referred to the clinic cohort before the end of study recruitment, that is, the data of a patient is subject to a univariate truncation $R \leqslant L$, where R is the logarithm of period from infection to referral to the liver clinic and L is the logarithm of time from infection to the

end of 1999. Time T, the logarithm of incubation period from disease infection to development of cirrhosis, is subject to right censoring at C, which is the logarithm of time from infection to last diagnosis follow-up. The purpose of the study was to determine how the progression to cirrhosis is affected by three covariates: age at infection, HIV-coinfection (yes:1 or no:0), and heavy alcohol consumption (yes:1 or no:0). A heavy alcohol intaker was defined as one consuming more than 50 units alcohol per week for at least five years.

Table 3 summarizes the estimates of regression parameters obtained from our method. The results from the truncated model, where the referral bias is considered, show that age at infection, HIV-coinfection and heavy alcohol in-take are significantly identified as risk factors associated with more rapid disease progression. If we compare the results with those from an untruncated model, ignorance of the referral bias has failed to identify heavy alcohol in-take as a significant risk factor. In medical literatures, older age at infection, HIV-coinfection and heavy alcohol intake have all been identified as factors associated with more rapid hepatitis C disease progression (Sharma and Sherker 2006). Based on our coefficient estimates shown in Table 3, we use $\hat{T} = \exp(W\beta)$ to predict the time period from infection to cirrhosis for individuals with different values of covariates. The prediction results and the corresponding values of covariates are shown in Figure 2. We can see that if we consider referral bias, the predicted values of X are larger than those without considering referral bias. This is because that the patients with more rapid disease progression are preferentially referred to the liver clinic cohort. Hence the incubation period from infection to cirrhosis observed in the clinic cohort may be shorter than that for the whole HCV patients community. If so, removing the referral bias, the predicted values should be larger comparing with the case only censoring is involved, as revealed by Fu *et al.* (2007).

Table 3 and Figure 1 are about here.

5. Discussion

In this paper, we considered the accelerated failure time model $\log T = \boldsymbol{W}\boldsymbol{\beta} + \varepsilon$ and a non-parametric method was proposed to allow a flexible bivariate distribution structure between T and another event time R. The dependency of the two event times and the referral bias in the sampling, incorporated by a right truncation on R, are both taken into account. We applied our proposed method to analyze the Edinburgh hepatitis C data in Fu *et al.* (2007). We concluded that when referral bias is considered, older age at HCV infection, HIV infection and heavy alcohol intake can all be identified as factors associated with more rapid HCV disease progression. Furthermore, the prediction results indicated that considering the referral bias will lead to longer estimate of time period from HCV infection to cirrhosis, which could reach an agreement with Fu *et al.* (2007). Our method can remove this kind of referral bias to get more reliable estimates for $\hat{\boldsymbol{\beta}}$. In this paper, our main interest is to study the effect of covariates on progression time to cirrhosis. The methodology could be extended to model the covariates' effect on referral time as well. We did not do this since it is of less interest in our cirrhosis study. If patients joined hospital after developing cirrhosis (this is unlikely and does not occur in our data set), diagnosis of cirrhosis is likely to be delayed and we may not know the exact time of T. For such extreme cases, T will hence be left censored by R and new models and estimating methods should be developed since dependent censoring is involved. We leave this as future work.

To find the estimate for $\boldsymbol{\beta}$ in model (1.1), we may also extend the Koul-Susarla-Van (KSV) methods by using the polar-coordinate transformation method. The KSV method is also based on an inverse probability weighted approach, but it only re-weights the response variable. In contrast, the weighted least squares re-weight both the response variable and the predictors. It is worth making further comparisons for these two methods under both censoring and truncation. The independent assumption between (C, L) and $\boldsymbol{W}$, required by the KSV or IPW estimator, may not be appropriate in other studies. It could be possible to adapt our method under a weaker

assumption, conditional independence of (C, L) and (R, T) given $\boldsymbol{W}$. We leave this to a future work.

An alternative to the AFT models is the proportional hazard model, which can be written as a transformation model (Chen *et al.*, 2002). The methods proposed here can be extended to proportional hazards models with time-independent covariates since the parameter estimation problem for the transformation model can also be solved using the inverse probability weighted method. In this paper we did not consider the proportional hazard model since the Schoenfeld residuals checking (Collett, 2003) implies that the proportional hazard assumption may be violated for the variable 'Heavy alcohol'. See Figure 2. We are currently working on extending the proposed method to proportional hazards models with time-varying coefficients.

Figure 2 is about here.

APPENDIX

A. Discussion on the boundary conditions

The condition (i) requires that the maximum values that R and T can take are smaller than those for random variables L and C respectively. This is reasonable in practice, because for any large values of R or T, it is always possible to collect data for an individual who was infected long time ago with larger L or C. Our basic assumption is that the observed data are from HCV-infected population who were or will be eventually referred to hospital. We can not use clinic cohort data to make inference for those who are never referred to hospital. Similarly, the condition (ii) guarantees that $G(t_1, t_2)$ is identifiable. It is also reasonable in practice. Even for a very small value of L, HCV-diagnosis and subsequent referral to liver clinics may happen shortly after infection because of regular HCV screening (more available in recent years), will give a very small value of $R \leqslant L$. For a very small value of C, it is possible to have a patient who developed cirrhosis right after referral, which will give a small value of $T \leqslant C$.

B. Proof of Theorem 3.1

The estimating equation (3.10) can be written as

$$\sqrt{n}\left[\boldsymbol{Q}\left(\boldsymbol{\beta};\hat{G}\right)\right]=\sqrt{n}\left[\boldsymbol{Q}\left(\boldsymbol{\beta};G\right)\right]+\sqrt{n}\left[\boldsymbol{Q}\left(\boldsymbol{\beta};\hat{G}\right)-\boldsymbol{Q}\left(\boldsymbol{\beta};G\right)\right]. \tag{B.1}$$

The first term of (B.1), $\sqrt{n}\left[\boldsymbol{Q}\left(\boldsymbol{\beta};G\right)\right]=n^{-1/2}\sum\limits_{i=1}^{n}\dfrac{\delta_i^*}{G(R_i^*,X_i^*)}(X_i^*-\boldsymbol{W}_i^*\boldsymbol{\beta})\boldsymbol{W}_i^{*tr}$ is a sum of i.i.d. variables. Now we focus on the second term of the right side of (B.1) which can be written as

$$\begin{aligned}
&\sqrt{n}\left[\boldsymbol{Q}\left(\boldsymbol{\beta};\hat{G}\right)-\boldsymbol{Q}\left(\boldsymbol{\beta};G\right)\right]\\
&=n^{-1/2}\sum_{i=1}^{n}\left[\frac{G(R_i^*,X_i^*)-\hat{G}(R_i^*,X_i^*)}{G^2(R_i^*,X_i^*)}\right]\boldsymbol{\eta}_i\\
&\quad+n^{-1/2}\sum_{i=1}^{n}\frac{G(R_i^*,X_i^*)-\hat{G}(R_i^*,X_i^*)}{G(R_i^*,X_i^*)}\left[\frac{1}{\hat{G}(R_i^*,X_i^*)}-\frac{1}{G(R_i^*,X_i^*)}\right]\boldsymbol{\eta}_i\\
&:=I+II.
\end{aligned} \tag{B.2}$$

Using the result (2.9) we have that as $n\to\infty$, II in (B.2) converges to 0 in probability. Hence,

$$\sqrt{n}\left[\boldsymbol{Q}\left(\boldsymbol{\beta};\hat{G}\right)-\boldsymbol{Q}\left(\boldsymbol{\beta};G\right)\right]=n^{-1/2}\sum_{i=1}^{n}\left[\frac{G(R_i^*,X_i^*)-\hat{G}(R_i^*,X_i^*)}{G^2(R_i^*,X_i^*)}\right]\boldsymbol{\eta}_i+o_p(1). \tag{B.3}$$

Based on the observed transformed data defined in (2.7) and the results in (2.9), equation(B.3) can be written as

$$\begin{aligned}
&\sqrt{n}\left[\boldsymbol{Q}\left(\boldsymbol{\beta};\hat{G}\right)-\boldsymbol{Q}\left(\boldsymbol{\beta};G\right)\right]=\sum_{i=1}^{n}\frac{n^{-3/2}\boldsymbol{\eta}_i}{G\left(\tilde{Z}_i^*(\alpha_i);\alpha_i\right)}\left[\sum_{j=1}^{n}\int_{s\leqslant\tilde{Z}_i^*(\alpha_i)}\frac{1}{H(s-;\alpha_i)}M_j(ds;\alpha_i)\right]+o_p(1)\\
&=n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\xi_{ji}\boldsymbol{\eta}_i}{G\left(\tilde{Z}_i^*(\alpha_i);\alpha_i\right)}+o_p(1):=\boldsymbol{U_n}+o_p(1).
\end{aligned} \tag{B.4}$$

According to the properties of U-statistics (Serfling, 1980) we have $\boldsymbol{U}_n=\widehat{\boldsymbol{U}}_{\boldsymbol{n}}+o(n^{-1}(\log n)^{\varrho})$ for some $\varrho>0$, where $\widehat{\boldsymbol{U}}_{\boldsymbol{n}}=\sum\limits_{k=1}^{n}E\left\{\boldsymbol{U_n}\,|\mathcal{D}_k\right\}=\sum\limits_{k=1}^{n}E\left\{n^{-3/2}\sum\limits_{i,j=1}^{n}G^{-1}(R_i^*,X_i^*)\xi_{ji}\boldsymbol{\eta}_i\,|\mathcal{D}_k\right\}$. Since

ξ_{ji}, defined in (3.11), is a zero-mean martingale with $\mathcal{D}_i$ given, we have $E\{\xi_{ji}|\mathcal{D}_i\} = \mathbf{0}$ (Dai and Fu, 2012). Thus if $j \neq k$, then $E\left\{G^{-1}(R_i^*, X_i^*)\xi_{ji}\boldsymbol{\eta}_i|\mathcal{D}_k\right\} = E\left\{G^{-1}(R_i^*, X_i^*)\xi_{ji}\boldsymbol{\eta}_i\right\} = \mathbf{0}$. So we have,

$$\begin{aligned}\widehat{\boldsymbol{U}}_{\boldsymbol{n}} &= n^{-1/2}\sum_{k=1}^{n} E\left\{G^{-1}(R_i^*, X_i^*)\xi_{ki}\boldsymbol{\eta}_i|\mathcal{D}_k\right\} + o_p(1)\\ &:= n^{-1/2}\sum_{k=1}^{n}\boldsymbol{\Phi}(\mathcal{D}_k) + o_p(1). \end{aligned} \tag{B.5}$$

The above equation together with (B.4) imply that $\sqrt{n}\left[\boldsymbol{Q}\left(\boldsymbol{\beta};\hat{G}\right) - \boldsymbol{Q}\left(\boldsymbol{\beta};G\right)\right] = n^{-1/2}\sum\limits_{k=1}^{n}\boldsymbol{\Phi}(\mathcal{D}_k) + o_p(1)$ which is a sum of i.i.d. terms.

Then the variance-covariance matrix of $\sqrt{n}\left[\boldsymbol{Q}\left(\boldsymbol{\beta};\hat{G}\right)\right]$ is given by

$$\boldsymbol{\Sigma}_{\boldsymbol{Q}} = \mathrm{Var}\left\{\sqrt{n}\left[\boldsymbol{Q}(\boldsymbol{\beta};G)\right] + \sqrt{n}\left[\boldsymbol{Q}(\boldsymbol{\beta};\hat{G}) - \boldsymbol{Q}(\boldsymbol{\beta};G)\right]\right\} = \mathrm{Var}\left[\frac{\boldsymbol{\eta}_i}{G(R_i^*, X_i^*)} + \boldsymbol{\Phi}(\mathcal{D}_i)\right].$$

Then the theorem follows from the first-order Taylor extension

$$\boldsymbol{Q}(\boldsymbol{\beta}_2^*;\hat{G}) = \boldsymbol{Q}(\hat{\boldsymbol{\beta}}_2;\hat{G}) + \boldsymbol{Q}^{'}(\hat{\boldsymbol{\beta}}_2;\hat{G})\left(\boldsymbol{\beta}_2^* - \hat{\boldsymbol{\beta}}_2\right) = \boldsymbol{Q}^{'}(\hat{\boldsymbol{\beta}}_2;\hat{G})\left(\boldsymbol{\beta}_2^* - \hat{\boldsymbol{\beta}}_2\right).$$

References

Akritas M.G. and Keilegom I.V. (2003). Estimation of bivariate and marginal distributions with censored data. *JRSS*, **B**(65): 457-471.

Bao Y., He S. and Mei C. (2007). The Koul-Susarla-Van Ryzin and weighted least squares estimates for censored linear regression model: A comparative study. *Computational Statistics & Data Analysis*, **51**: 6488-6497.

Buckley J. and James I. (1979). Linear regression with censored data. *Biometrika*, **66**: 429-436.

Burke M.D. (1988). Estimation of a bivariate distribution function under random censorship.

Biometrika, **75**: 379-382.

Campbell G. (1981). Nonparametric bivariate estimation with randomly censored data. *Biometrika*, **68**: 417-422.

Chen K., Jin Z. and Ying Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, **89**: 659-668.

Clayton D.G. (1978). A model for association in bivariate life tables and its application in epidemiology studies of familial tendency in chronic disease incidence. *Biometrika*, **65**: 141-151.

Collett D. (2003). *Modelling Survival Data in Medical Research*, Chapman & Hall, CRC.

Dabrowska D.M. (1988). Kaplan-Meier estimate on the plane. *Annals of Statistics*, **16**: 1475-1489.

Dai H. and Bao Y. (2009). An inverse probability weighted estimator for the bivariate distribution function under right censoring. *Statistics and Probability Letters*, **79**: 1789-1797.

Dai H. and Fu B. (2012). A polar coordinate transformation for estimating bivariate survival functions with randomly censored and truncated data. *J. of Statistical Planning and Inference*, **142**:248-262.

Dore G.J., Freeman A.J., Law M., and Kaldor M. (2002). Advances in Liver Disease: Hepatitis C – Is severe liver disease a common outcome for people with chronic hepatitis C? *J. of Gastroenterology and Hepatology*, **17**: 423-430.

Freeman A.J., Dore G.J., Law M.G., Thorpe M., Von Overbeck J. and Lloyd A.R., et al. (2001). Estimating progression to cirrhosis in chronic hepatitis C virus infection. *Hepatology*, **34**: 809-816.

Fu B., Tom B., Delahooke T., Alexander G.J.M. and Bird S.M. (2007) Event-biased referral can distort estimation of hepatitis C virus progression rate to cirrhosis and of prognostic influences. *J. of Clinical Epidemiology*, **60**: 1140-1148.

Gijbels I. and Gurler U. (1998). Covariance function of a bivariate distribution function estimator for left truncated and right censored data. *Statistica Sinica*, 1219-1232.

Gross S. T. and Huber-Carol C. (1992). Regression models for truncated survival data. *Scand. J. Statist*, 19, 193-213.

Gross S. T. and Lai T. L. (1996). Nonparametric estimation and regression analysis with left-truncated and right-censored data. *J. Amer. Statist. Assoc.*, 91, 1166-1180.

Gurler U. (1996). Bivariate estimation with right truncated data. *J. of Amer. Statist. Assoc.*, **91**: 1152-1165.

Gurler U. (1997). Bivariate distribution and hazard functions when a component is randomly truncated. *J. of Multivariate Analysis*, **60**: 20-47.

He S. and Yang G. (2003). Estimation of regression parameters with left truncated data. *J. of Statistical Planning and Inference*, **117**: 99-122.

He S. and Wong X. (2003). The central limit theorem of linear regression model under right censorship. *Science in China Series A*, **46**: 600-610.

Huang J., Vieland V.J. and Wang K. (2001). Nonparametric estimation of marginal distributions under bivariate truncation with application to testing for age-of-Onset anticipation. *Statistica Sinica*, **11**: 1047-1068.

Jin Z., Lin D. Y. and Ying Z. (2006). On least-squares regression with censored data. *Biometrika*, **93**: 147-161.

Koul H., Susarla V. and Van Ryzin J. (1981). Regression analysis with randomly right censored data. *The Annals of Statistics*, **9**: 1276-1288.

Lin D.Y. and Ying Z. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika*, **80**(3): 572-581.

Miller R. (1976). Least square regression with censored data. *Biometrika*, **63**: 449-464.

Miller R. and Halpern J. (1982). Regression with censored data. *Biometrika*, **69**: 521-531.

Prentice R.L., Moodie F.Z. and Wu J. (2004). Hazard-based nonparametric survivor function estimation. *JRSS*, **B**(66): 305-319.

Serfling J. Robert (1980). Approximation Theorems of Mathematical Statistics. *John Wiley & Sons*, Inc.

Sharma N.K. and Sherker A.H. (2010). Epidemiology, Risk Factors, and Natural History of Chronic Hepatitis C. *Clinical Gastroenterology, 33-70*, edited by K. Shetty and G.Y. Wu. Humana Press.

Shen P. (2006). An inverse-probability-weighted approach to estimation of the bivariate survival function under left-truncation and right-censoring. *J. of Statist. Planning and Inference*, **136**: 4365-4384.

Tsai W.Y. and Leurgan S. and Crowley J. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring. *The Annals of Statistics*, **14**: 1351-1365.

van der Laan M.J. (1996a). Efficient estimation in the bivariate censoring model and repairing NPMLE. *The Annals of Statistics*, **24**: 596-627.

van der Laan M.J. (1996b). Nonparametric estimation of the bivariate survival function with truncated data. *J. of Multivariate Analysis*, **58**: 107-131.

Table 1. Scenario 1. (i) true values; (e) estimate (bias in parenthesis). n: observed sample size; γ: truncation probability; $\%c$: censoring percentage; $\hat{s}_\beta$: means of standard deviation estimates obtained by (3.13); $\hat{\sigma}_\beta$: the standard deviation for $\hat{\beta}$ and $\tilde{\beta}$ based on the 500 simulations. (1) $c = d = 1.1$ corresponding to truncation probability $\gamma \approx 0.85$ and $a = b = 1.8$, $a = b = 1.0$ and $a = b = 0.65$ approximately corresponding to 15% censoring, 50% censoring and 80% censoring, respectively; (2) $c = d = 0.45$ corresponding to $\gamma \approx 0.5$ and $a = b = 1.3$, $a = b = 0.8$ and $a = b = 0.55$ approximately corresponding to 15% censoring, 50% censoring and 80% censoring, respectively; (2) $c = d = 0.2$ corresponding to $\gamma \approx 0.15$ and $a = b = 1.1$, $a = b = 0.7$ and $a = b = 0.45$ approximately corresponding to 15% censoring, 50% censoring and 80% censoring, respectively.

$n = 200$	c% = 20%			c% = 50%			c% = 80%		
$\gamma =$	(e)	$\hat{s}_\beta$	$\hat{\sigma}_\beta$	(e)	$\hat{s}_\beta$	$\hat{\sigma}_\beta$	(e)	$\hat{s}_\beta$	$\hat{\sigma}_\beta$
0.85	3.699(0.001)	0.115	0.112	3.717(0.017)	0.160	0.156	3.749(0.049)	0.269	0.246
0.85	-0.051(0.001)	0.004	0.004	-0.051(0.001)	0.006	0.006	-0.051(0.001)	0.010	0.009
0.85	-0.301(0.001)	0.024	0.025	-0.305(0.005)	0.035	0.035	-0.304(0.004)	0.068	0.059
0.85	-0.099(0.001)	0.025	0.025	-0.103(0.003)	0.035	0.034	-0.102(0.002)	0.058	0.055
0.5	3.708(0.008)	0.126	0.131	3.748(0.048)	0.176	0.171	3.741(0.041)	0.317	0.272
0.5	-0.050(0.000)	0.005	0.005	-0.051(0.001)	0.007	0.007	-0.051(0.001)	0.012	0.010
0.5	-0.305(0.005)	0.029	0.028	-0.304(0.004)	0.038	0.037	-0.308(0.008)	0.068	0.058
0.5	-0.100(0.000)	0.029	0.028	-0.101(0.001)	0.036	0.035	-0.103(0.003)	0.058	0.055
0.15	3.649(0.051)	0.217	0.183	3.755(0.055)	0.255	0.198	3.759(0.059)	0.363	0.278
0.15	-0.049(0.001)	0.008	0.007	-0.050(0.001)	0.010	0.008	-0.051(0.001)	0.014	0.010
0.15	-0.287(0.013)	0.054	0.043	-0.298(0.002)	0.056	0.044	-0.290(0.010)	0.080	0.058
0.15	-0.099(0.001)	0.058	0.043	-0.102(0.002)	0.056	0.043	-0.096(0.004)	0.076	0.055

$n = 100$	c% = 20%			c% = 50%			c% = 80%		
$\gamma =$	(e)	$\hat{s}_\beta$	$\hat{\sigma}_\beta$	(e)	$\hat{s}_\beta$	$\hat{\sigma}_\beta$	(e)	$\hat{s}_\beta$	$\hat{\sigma}_\beta$
0.85	3.725(0.025)	0.162	0.161	3.726(0.026)	0.241	0.216	3.665(0.335)	0.446	0.325
0.85	-0.051(0.001)	0.006	0.006	-0.050(0.000)	0.009	0.008	-0.048(0.002)	0.017	0.012
0.85	-0.301(0.001)	0.039	0.037	-0.302(0.002)	0.055	0.051	-0.295(0.005)	0.102	0.073
0.85	-0.100(0.000)	0.037	0.035	-0.100(0.000)	0.049	0.048	-0.107(0.007)	0.094	0.071
0.5	3.710(0.010)	0.180	0.180	3.664(0.036)	0.250	0.229	3.726(0.026)	0.429	0.338
0.5	-0.051(0.001)	0.007	0.007	-0.052(0.002)	0.009	0.009	-0.051(0.001)	0.015	0.012
0.5	-0.304(0.004)	0.041	0.040	-0.308(0.008)	0.053	0.050	-0.292(0.008)	0.099	0.069
0.5	-0.098(0.002)	0.040	0.040	-0.102(0.002)	0.053	0.049	-0.101(0.001)	0.091	0.069
0.15	3.638(0.062)	0.266	0.202	3.655(0.045)	0.320	0.222	3.764(0.064)	0.528	0.343
0.15	-0.049(0.001)	0.010	0.008	-0.048(0.002)	0.012	0.009	-0.049(0.001)	0.020	0.013
0.15	-0.293(0.007)	0.063	0.048	-0.295(0.005)	0.067	0.050	-0.285(0.015)	0.109	0.067
0.15	-0.099(0.001)	0.067	0.047	-0.091(0.009)	0.069	0.051	-0.094(0.006)	0.097	0.062

Table 2. (i) true values; (e) estimate (bias in parenthesis). n: observed sample size; γ: truncation probability; $\%c$: censoring percentage; $\hat{s}_\beta$: means of standard deviation estimates obtained by (3.13); $\hat{\sigma}_\beta$: the standard deviation for $\hat{\beta}$ and $\hat{\beta}$ based on the 500 simulations. **Scenario 2.** (1) $c = d = 0.9$ corresponding to truncation probability $\gamma \approx 0.8$ and $a = b = 1.5$, $a = b = 1.0$ and $a = b = 0.7$ approximately corresponding to 25% censoring, 50% censoring and 75% censoring, respectively; (2) $c = d = 0.5$ corresponding to $\gamma \approx 0.5$ and $a = b = 1.2$, $a = b = 0.9$ and $a = b = 0.6$ approximately corresponding to 25% censoring, 50% censoring and 75% censoring, respectively; (3) $c = d = 0.2$ corresponding to $\gamma \approx 0.25$ and $a = b = 1.0$, $a = b = 0.7$ and $a = b = 0.5$ approximately corresponding to 25% censoring, 50% censoring and 75% censoring, respectively. **Scenario 3.** (1) $c = d = 0.8$ corresponding to $\gamma \approx 0.85$ and $a = b = 1.9$, $a = b = 1.4$ and $a = b = 1.0$ approximately corresponding to 25% censoring, 50% censoring and 75% censoring, respectively; (2) $c = d = 0.35$ corresponding to $\gamma \approx 0.5$ and $a = b = 1.8$, $a = b = 1.2$ and $a = b = 0.85$ approximately corresponding to 25% censoring, 50% censoring and 75% censoring, respectively.

				Scenario 2					
$n = 200$	c% = 25%			c% = 50%			c% = 75%		
$\gamma =$	(e)	$\hat{s}_\beta$	$\hat{\sigma}_\beta$	(e)	$\hat{s}_\beta$	$\hat{\sigma}_\beta$	(e)	$\hat{s}_\beta$	$\hat{\sigma}_\beta$
0.8	3.665(0.005)	0.220	0.208	3.621(0.079)	0.307	0.286	3.573(0.137)	0.427	0.377
0.8	-0.049(0.001)	0.008	0.008	-0.049(0.001)	0.012	0.011	-0.048(0.002)	0.016	0.015
0.8	-0.298(0.002)	0.048	0.047	-0.289(0.011)	0.062	0.061	-0.279(0.021)	0.096	0.087
0.8	-0.099(0.001)	0.047	0.046	-0.102(0.002)	0.063	0.061	-0.093(0.007)	0.095	0.086
0.5	3.646(0.054)	0.218	0.212	3.637(0.063)	0.262	0.248	3.561(0.139)	0.430	0.382
0.5	-0.050(0.000)	0.008	0.008	-0.049(0.001)	0.010	0.010	-0.048(0.002)	0.016	0.015
0.5	-0.290(0.010)	0.047	0.048	-0.296(0.004)	0.060	0.060	-0.274(0.026)	0.010	0.090
0.5	-0.096(0.004)	0.049	0.048	-0.105(0.005)	0.060	0.058	-0.095(0.005)	0.095	0.088
0.2	3.590(0.110)	0.319	0.296	3.547(0.153)	0.346	0.316	3.514(0.186)	0.441	0.391
0.2	-0.048(0.002)	0.013	0.012	-0.047(0.003)	0.014	0.013	-0.046(0.004)	0.017	0.015
0.2	-0.289(0.011)	0.070	0.068	-0.278(0.022)	0.080	0.076	-0.265(0.035)	0.111	0.089
0.2	-0.093(0.007)	0.073	0.068	-0.094(0.006)	0.082	0.074	-0.098(0.002)	0.105	0.089

				Scenario 3					
$n = 200$	c% = 25%			c% = 50%			c% = 75%		
$\gamma =$	(e)	$\hat{s}_\beta$	$\hat{\sigma}_\beta$	(e)	$\hat{s}_\beta$	$\hat{\sigma}_\beta$	(e)	$\hat{s}_\beta$	$\hat{\sigma}_\beta$
0.85	3.409(0.291)	0.520	0.512	3.132(0.568)	0.723	0.642	2.521(1.179)	0.971	0.745
0.85	-0.042(0.008)	0.020	0.020	-0.033(0.017)	0.027	0.025	-0.019(0.031)	0.038	0.029
0.85	-0.253(0.047)	0.126	0.117	-0.202(0.098)	0.155	0.137	-0.071(0.229)	0.245	0.180
0.85	-0.082(0.018)	0.120	0.117	-0.076(0.024)	0.150	0.138	-0.064(0.036)	0.234	0.177
0.5	2.878(0.822)	1.270	0.845	2.833(0.867)	1.400	0.970	2.272(1.418)	1.683	1.140
0.5	-0.041(0.009)	0.049	0.033	-0.043(0.007)	0.058	0.037	-0.024(0.026)	0.065	0.044
0.5	-0.266(0.034)	0.295	0.203	-0.289(0.011)	0.348	0.214	-0.175(0.125)	0.373	0.239
0.5	-0.103(0.003)	0.281	0.204	-0.108(0.008)	0.325	0.213	-0.066(0.034)	0.368	0.246

Table 3. Demographic details for the Edinburgh liver clinic series and model estimation results.

Patient characteristics	
Mean age at HCV-infection (Age)	22.4
Number (%) of patients with known HIV status (HIV)	
Positive - 1	41 (11%)
Negative - 0	346 (89%)
Number (%) with heavy alcohol consumption (Alcohol)	
Yes - 1	116 (30%)
No - 0	271 (70%)

Estimation results (SE in parenthesis), * means the estimate is significant at 5% level.

	With truncation – $\hat{\boldsymbol{\beta}}$ (SE)	Without truncation – $\hat{\boldsymbol{\beta}}$ (SE)
Intercept	3.628* (0.068)	3.844* (0.133)
Age	-0.011* (0.003)	-0.031* (0.004)
HIV	-0.313* (0.047)	-0.380* (0.089)
Alcohol	-0.098* (0.038)	-0.077 (0.070)

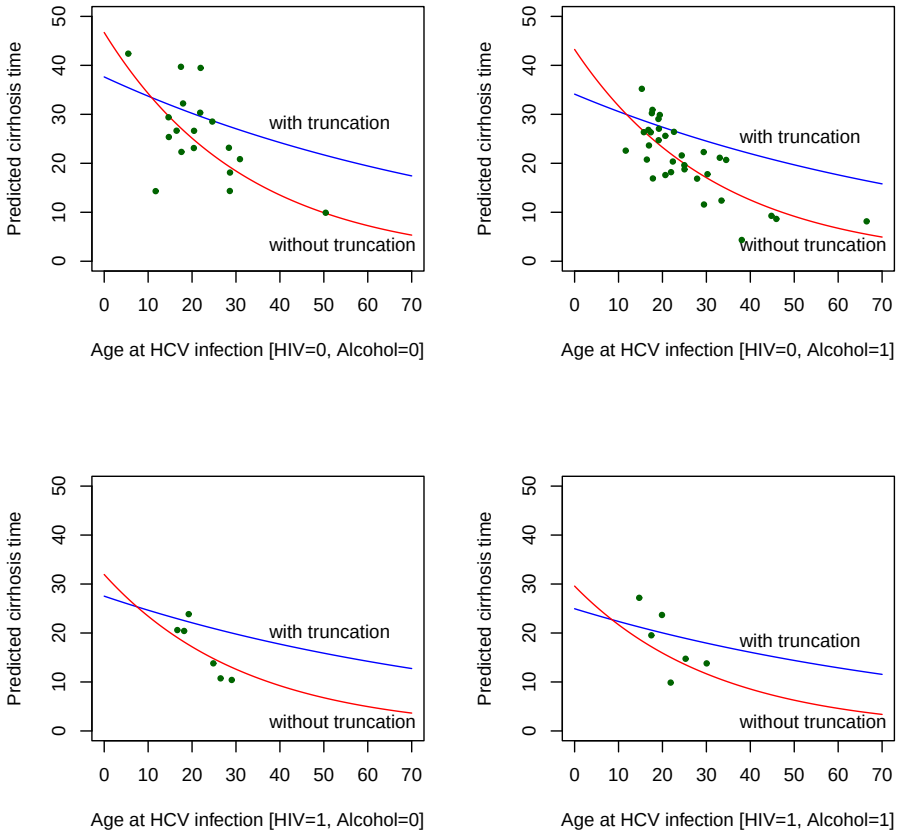

Fig. 1. Prediction for cirrhosis time.

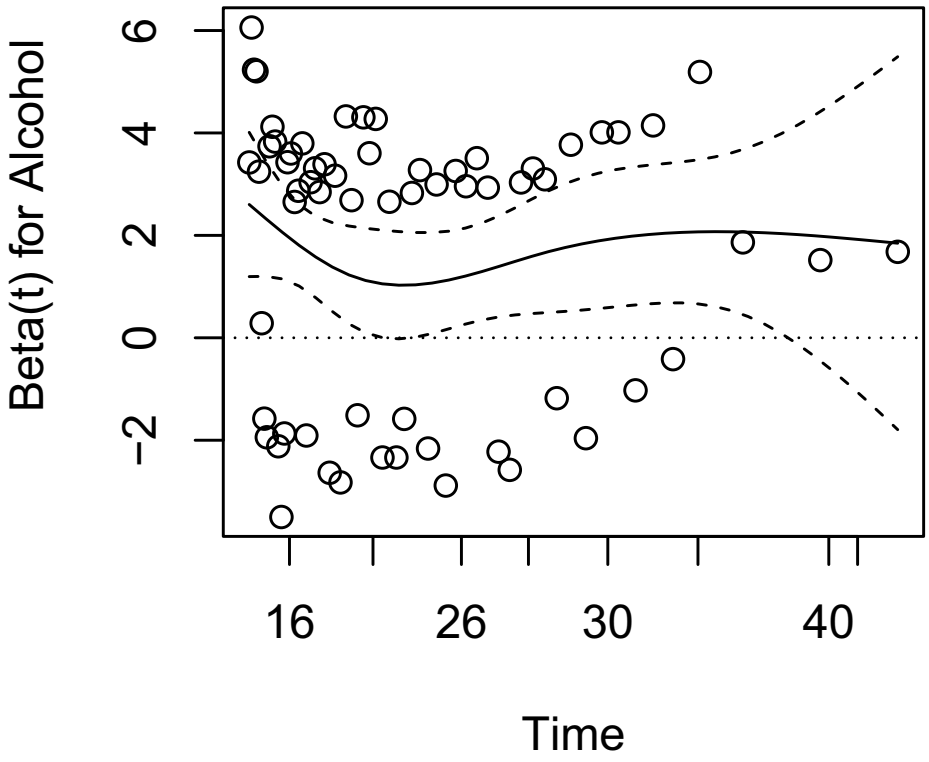

Fig. 2. Use of Schoenfeld residuals. It implies the parameter for Alcohol is not constant overtime.